smalldisco, a pipeline for siRNA discovery and 3' tail identification

Ian V. Caldas[1], Leanne H. Kelley[2], Yasir H. Ahmed-Braimah[2*], and Eleanor M. Maine[2*]

[1]Department of Molecular Biology and Genetics, Cornell University, 526 Campus Rd, Ithaca, NY 14850

[2]Department of Biology, Syracuse University, 107 College Place, Syracuse, NY 13210

*Corresponding authors: Email: yahmed@syr.edu (Y.H.A.); emmaine@syr.edu (E.M.M.)

Abstract

Capturing and sequencing small RNAs is standard practice, however identification of a group of these small RNAs—small interfering RNAs (siRNAs)—has historically been an inconsistent and vague practice. We present smalldisco, a command-line tool for small interfering RNA discovery and annotation from small RNA-seq datasets. smalldisco can distinguish short reads that map antisense to exons, termed siRNAs, annotate these siRNAs by genomic feature (e.g., gene or exon) defined by the user, and quantify their abundance. smalldisco also uses the program Tailor to quantify 3' non-templated nucleotides of siRNAs or any small RNA species. smalldisco is available for download from GitHub (https://github.com/ianvcaldas/smalldisco).

Short running head: smalldisco program for siRNA discovery

Keywords: small RNA, siRNA identification, small RNA 3' tailing modification

Introduction

Small, non-coding RNAs (sRNAs) are critical components of transcriptional and post-transcriptional gene regulatory mechanisms in plants, animals, and fungi (Castel and Martienssen 2013, Holoch and Moazed 2015, Ozata *et al.* 2019, Galagali and Kim 2020,

Ketting and Cochella 2021). Three prominent types of sRNAs — micro (mi) RNAs, PIWI-associated (pi) RNAs, and small interfering (si) RNAs — associate with Argonaute proteins and provide specificity via complementary base pairing with target RNAs (Iwakawa and Tomari 2022). MiRNAs and piRNAs are encoded in the genome, whereas siRNAs are generated from RNA templates such that siRNAs themselves are antisense to their genome origin. While miRNAs and piRNAs are identified as individual, encoded sequences, siRNAs map to variable positions across the template RNA, and siRNAs derived from any given RNA template may differ between individuals. It is therefore relatively straightforward to identify miRNA and piRNA sequences in sRNA-seq data sets by mapping reads to their canonical loci in the genome. In contrast, siRNA sequences are not straightforward to identify.

There have been two main approaches to identifying siRNAs: (i) computationally predicting siRNAs from sRNA sequencing data (e.g., Montgomery *et al.* 2012, Wei *et al.* 2019, Rogers and Phillips 2020, Scheer *et al.* 2021), or (ii) identifying subsets of siRNAs by immunoprecipitating a specific Argonaute protein and recovering associated sRNAs (e.g., Claycomb *et al.* 2009, Chaves *et al.* 2021, de Albuquerque *et al.* 2015, Svendsen *et al.* 2019, Xu *et al.*, 2018). Unfortunately, computational approaches used by most published studies to identify siRNAs remain cryptic and difficult to apply broadly. The sRNA immunoprecipitation approach has been very useful for linking individual Argonaute proteins to regulation of certain genes via specific siRNA guides. However, it does not provide a comprehensive view of all siRNA sequences in a transcriptome or of their abundance.

We sought to develop a method for identifying siRNAs within full sRNA-seq data sets regardless of their Argonaute association. Such an unbiased approach would be broadly applicable and useful for addressing various biological questions. We developed smalldisco, a <u>small</u>-interfering RNA <u>disco</u>very program, for parsing sRNA-seq data sets to identify siRNAs that are antisense to coding regions and hence target protein-coding genes. In addition, to provide information about sRNA 3' tailing, smalldisco can feed its discovered siRNAs into the Tailor program (Chou *et al.* 2015) to identify 3' non-templated nucleotide additions to these siRNA sequences and produce output in a user-friendly format as tables of siRNA targets, their tail modifications, and their abundances. While

smalldisco does not identify miRNAs or piRNAs (programs available for those purposes include unitas [Gebert et al. 2017], miRDeep* [An et al. 2013], and piPipes [Han et al. 2015]), it does provide the option of identifying 3' tailing modifications to miRNA and piRNA sequences from the same sRNA dataset.

Methods

smalldisco users will need the following files: FASTQ or BAM files for their samples of sRNA-seq data; a reference genome (FASTA); and an annotated (GTF or GFF) file of the genome of interest. A summary of the pipeline is diagrammed in Figure 1. smalldisco can be run in three modes: `align`, `sirna`, and `tail` (or `run_all`).

In the `align` mode, if BAM files are not provided, smalldisco will use HISAT2 (Kim *et al.* 2019; https://github.com/DaehwanKimLab/hisat2) to build a genome index and map reads (using default parameters) against the provided reference genome, resulting in a BAM file. The next step implements our definition of an siRNA: reads that map antisense to an exon—or a user-defined genomic feature—are kept for further processing. For this analysis, exons are obtained from the genome annotation. For genes with multiple isoforms, all annotated coding sequences that overlap on the same strand are merged. In addition, any regions where both strands are coding (two genes overlap) are removed from the analysis because we cannot definitively pinpoint that an sRNA is antisense to one coding strand or a remnant of an mRNA produced from the opposite strand. (In the *C. elegans* genome, ~1.1% of exons have some overlap, much of which is non-coding.)

The `sirna` mode implements smalldisco's siRNA discovery. Groups of overlapping sRNA reads that map antisense to exons are merged into a putative siRNA region if they fulfill two criteria: if there are more than *k* overlapping reads and if the regions span *n* or more base pairs. Both parameters are set by the user, with default *k=10* and *n=10.* The result is a BED file of putative siRNA regions. The "siRNA region" is the unit to which the user wishes to annotate siRNAs, e.g., whole genes or individual exons.

`tail` mode, the last and optional step of the pipeline, can quantify read tails for the sRNAs. A BED file of sRNA regions, which can be the list of putative siRNA regions generated by the previous step or any other predefined list, is used as input to Tailor

(https://github.com/jhhung/Tailor). Tailor maps the reads to the reference genome and adds a tag at the end of each read's alignment record to indicate the length and sequence of a tail, or to indicate the absence of a tail, producing a tagged BAM file. Tailor implements a default of zero mismatches in the templated sRNA sequence before searching for non-templated tails. Tagged reads are then mapped to the provided sRNA regions file and the tail tags are converted to nucleotides. The output is a CSV file of sRNA targets and tails. In addition to "untailed" and the tail types in the table, one row per target will include a total for "all_tails." The sum of "untailed" and "all_tails" equals the total number of reads for that target. If the user is not interested in tail quantification, the tail mode can still be run to obtain counts for total siRNA reads.

We validated smalldisco using publicly available datasets published by Xu *et al.* (2018) (PRJNA447866), Claycomb *et al.* (2009) (PRJNA123485), Scheer *et al*. (2021) (PRJNA624174), and Davis *et al*. (2018) (PRJNA398116). Samples used for each analysis are listed in Table S1. To analyze these datasets, FASTQ files were obtained from NCBI's Gene Expression Omnibus, trimmed of 3' adapters using cutadapt, and run through smalldisco under the `run_all` mode with default parameters. For Xu *et al.* and Claycomb *et al.* datasets, the smalldisco siRNA output was normalized to counts per million (CPM) and filtered to retain only siRNA-associated genes with $\geq$ 25 reads per million, the threshold used by these authors. To normalize our siRNA library as executed by Scheer *et al.*, we created a library by counting reads that map to exons, combined this library with the smalldisco siRNA output, and normalized to CPM. Finally, to benchmark smalldisco's performance against another sRNA discovery tool, segmentSeq (Hardcastle *et al.* 2011), we analyzed two samples from Davis *et al*. (2018), each with three replicates (SRR5931601, SRR5931602, SRR5931603, SRR5931604, SRR5931605, SRR5931606), and compared runtime, number of identified siRNA "loci," and resolution of siRNA region sizes.

Results and discussion

We validated the siRNA discovery function of smalldisco using publicly available datasets from two studies where *C. elegans* Argonaute-specific small RNAs were recovered by immunoprecipitation. Xu *et al.* (2018) immunoprecipitated WAGO-4 Argonaute and

recovered siRNAs (present at or above 25 CPM) that corresponded to 4756 target genes. smalldisco analysis of their immunoprecipitated sRNA dataset identified siRNAs corresponding to 4706 target genes using a 25 CPM cut off, including ~93% of those identified by Xu *et al.* (4413) (Figure 2A). Claycomb *et al*. (2009) immunoprecipitated CSR-1 Argonaute and identified siRNAs corresponding to 4147 target genes present at or above 25 CPM and enriched ≥ two-fold in the IP compared to input RNA. smalldisco identified 4209 siRNA target genes in this dataset, including ~81% of those reported by Claycomb *et al.* (2009) (Figure 2B). For both datasets, some sequences were identified as siRNAs only by smalldisco or only by the authors, and we speculate these reflect differences in filtering criteria or data processing. Overall, our results demonstrate that smalldisco can accurately identify siRNAs with known Argonaute association.

We validated the siRNA counting function using publicly available datasets from Scheer *et al.* (2021). This study sequenced sRNAs from wildtype and three mutant strains of *Arabidopsis* and identified siRNAs within those four datasets. The authors provide a brief description of how they identified siRNA reads wherein their method seems to use logic similar to ours except that siRNAs were not required to map antisense to coding regions. We evaluated the four datasets with smalldisco and compared the output to the authors' results. For each genotype, the number of siRNA reads identified by smalldisco was similar to that identified by the authors (Figure 3). Compared to the authors' analysis, smalldisco identified: 5% more siRNA reads in wildtype, ~the same number of reads in the *urt1-1* single mutant, 88% as many reads in the *xrn4-3* single mutant, and 85% as many reads in the *urt1-1 xrn4-3* double mutant.

One published protocol for identifying siRNAs and for which software is available is segmentSeq (Hardcastle *et al.* 2011). segmentSeq assumes all siRNAs are generated by Dicer endonuclease cleavage of double-stranded RNA, and therefore reads matching the sense strand should be present in the dataset. Although most *C. elegans* siRNAs are generated as direct single-strand RNA products of RNA-directed RNA polymerase and do not require Dicer activity, we were interested in comparing sRNA discovery by the two programs. We analyzed the Davis *et al.* (2018) dataset (see Methods) with each program and noted several major differences in output. First, smalldisco analyzed our sample input dataset in ~2.5 hr, whereas segmentSeq required ~2.5 weeks. Second, the two programs

identified a common set of 6830 sRNAs and another substantial number of sRNAs unique to one or the other program (1835 for segmentSeq, 3899 for smalldisco) (Figure 4A). Third, the two programs mapped sRNAs to the genome - and hence identified siRNA-associated genes - at different spatial scales. smalldisco identified an average siRNA region of 124.3 bp (± 7.4 bp SE) when run with the default "CDS" (exon) setting, whereas segmentSeq identified an average siRNA region of 6.7kb (± 68.9 bp SE) corresponding, in some cases, to more than one gene (Figure 4B, Figure S1). Although segmentSeq is a useful tool for siRNA identification, smalldisco is more complete, gives finer grained results, and is faster.

In summary, smalldisco is a command-line tool for siRNA discovery and annotation from small RNA sequencing datasets. We expect that smalldisco will be useful in identifying, annotating, and quantifying the abundance of siRNAs in wildtype and mutant datasets derived from whole animals and dissected tissue. Inclusion of the Tailor program in the pipeline allows the user to identify and quantify 3' non-templated nucleotides that may be present on the siRNAs or any other small RNA in the dataset.

Data availability

smalldisco and supporting documentation can be downloaded from GitHub (https://github.com/ianvcaldas/smalldisco).

Acknowledgements

The authors thank Sarah Hall for useful comments on the manuscript.

Funding

LHK was supported by a Syracuse University Graduate Fellowship. Funding was also provided by the National Institutes of Health (grant R03HD0091645) to EMM.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

An J, Lai J, Lehman ML, Nelson CC. 2013. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res 41: 27-737.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Rev Genet 14:100-112.

Chaves DA, Dai H, Li L, Moresco JJ, Oh ME, Conte Jr D, Yates III JR, Mello CC, Gu W. 2021. The RNA phosphatase PIR-1 regulates endogenous small RNA pathways in *C. elegans*. Mol Cell 81:1-12.

Chou MT, Han BW, Hsiao CP, Zamore PD, Weng Z, Hung JH. 2015. Tailor: a computational framework for detecting non-templated tailing of small silencing RNAs. NAR 43(17):e109. doi: 10.1093/nar/gkv537.

Claycomb JM, Batista PJ, Pang KM, Gu W, Vasale JJ, van Wolfswinkel JC, Chaves DA, Shirayama M, Mitani S, Ketting RF, *et al*. 2009. The Argonaute CSR-1 and its 22G-RNA cofactors are required for homocentric chromosome segregation. Cell 139:123-134.

Davis GM, Shikui T, Anderson JWT, Colson RN, Gunzburg MJ, Francisco MA, Ray D, Shrubsole SP, Sobotka JA, Seroussi U, Lao RX, Maity T, Su MZ, *et al.* 2018. The TRIM-NHL protein NHL-2 is a co-factor in the nuclear and somatic RNAi pathways in *C. elegans*. eLife 7:e35478. doi: 10.7554/eLife.35478.

de Albuquerque BFM, Placentino M, Ketting RF. 2015. Maternal piRNAs are essential for germline development following de novo establishment of endo-siRNAs in *Caenorhabditis elegans*. Dev Cell 34:228-256.

Galagali H, Kim JK. 2020. The multifaceted roles of microRNAs in differentiation. Curr Op Cell Biol 67:118-140.

Gebert D, Hewel C, Rosenkranz D. 2017. unitas: the universal tool for annotation of small RNAs. BMC Genet 18: 1-14.

Han BW, Wang W, Zamore PD, Weng Z. 2015. piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics 31:593-595.

Hardcastle TJ, Kelly KA, Baulcombe DC. 2011. Identifying small interfering RNA loci from high-throughput sequencing data. Bioinformatics 28:457-463.

Holoch D, Moazed D. 2015. RNA-mediated epigenetic regulation of gene expression. Nature Rev Genet 16:71-84.

Iwakawa H, Tomari Y. 2022. Life of RISC: formation, action, and degradation of RNA-induced silencing complex. Mol Cell 82: 31-43. doi: 10.1016/j.molcel.2021.11.026.

Ketting RF, Cochella L. 2021. Concepts and functions of small RNA pathways in *C. elegans*. Curr Top Dev Biol 144:45-89.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotech 37:907-915.

Montgomery TA, Rim YS, Zhang C, Dowen RH, Phillips CM, Fischer SEJ, Ruvkun G. 2012. PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog *henn-1*. PLoS Genet. 8:1-12.

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. Nature Rev Genet 20:89-108.

Rogers AR, Phillips CM. 2020. A small-RNA-mediated feedback loop maintains proper levels of 22G-RNAs in *C. elegans*. Cell Reports 33(3):1-13.

Scheer H, de Almeida C, Ferrier E, Simonnot Q, Poirier L, Pflieger D, Sement FM, Koechler S, Piermaria C, Krawczyk P, Mrocsek S, *et al.* 2021. The TUTase URT1 connects decapping activators and prevents the accumulation of excessively deadenylated mRNAs to avoid siRNA biogenesis. Nat Commun 12:1298. doi: 10.1038/s41467-021-21382-2.

Svendsen JM, Reed KJ, Vijayasarathy T, Montgomery BE, Tucci RM, Brown KC, Marks TN, Nguyen DAH, Phillips CM, Montgomery TA. 2019. Henn-1/HEN1 promotes germline immortality in *Caenorhabditis elegans*. Cell Rep 29:3187-3199.e4.

Wei Y, Li F, Zhang S, Zhang S, Zhang H, Sun R. 2019. Analysis of small RNA changes in different *Brassica napus* synthetic allopolyploids. PeerJ 7:e7621 doi: 10.7717/peerj.7621

Xu R, Feng X, Chen X, Weng C, Yan Q, Xu T, Hong M, Guang S. 2018. A cytoplasmic Argonaute protein promotes the inheritance of RNAi. Cell Rep 23:2482-2494.

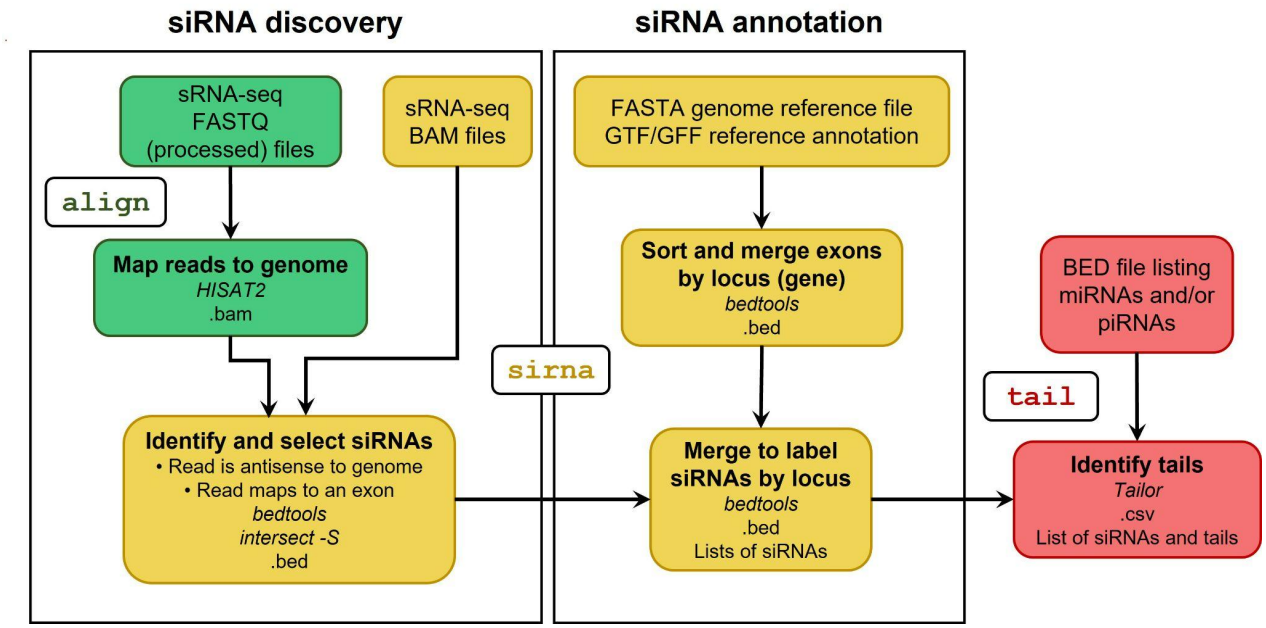

Figure 1 Overview of the smalldisco pipeline outlining the siRNA discovery and annotation workflows. Three analysis modes are available: `align`, `sirna`, and `tail`.

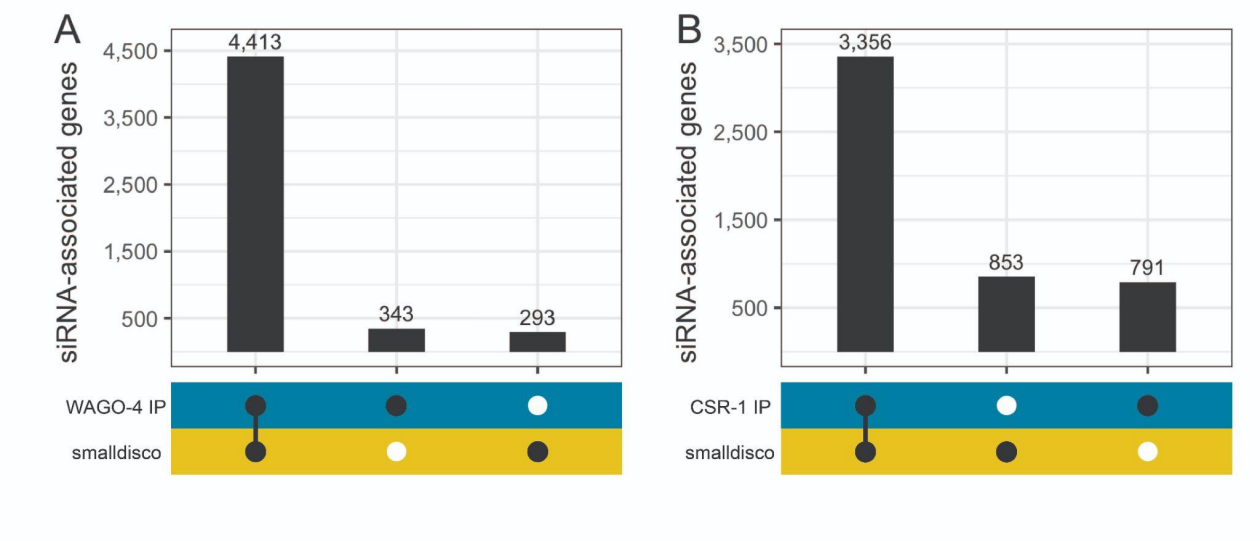

Figure 2 siRNA identification among sRNAs that co-immunoprecipitate with Argonaute proteins. Small RNAs enriched in (A) WAGO-4 immunoprecipitate (Xu *et al.* 2018) and (B) CSR-1 immunoprecipitate (Claycomb *et al.* 2009) were analyzed with smalldisco and the output was compared to the authors' reported findings. Plots indicate the number of siRNA target genes identified in the immunoprecipitated sRNA dataset by the authors' analysis (WAGO-4 IP, CSR-1 IP) and/or by smalldisco.

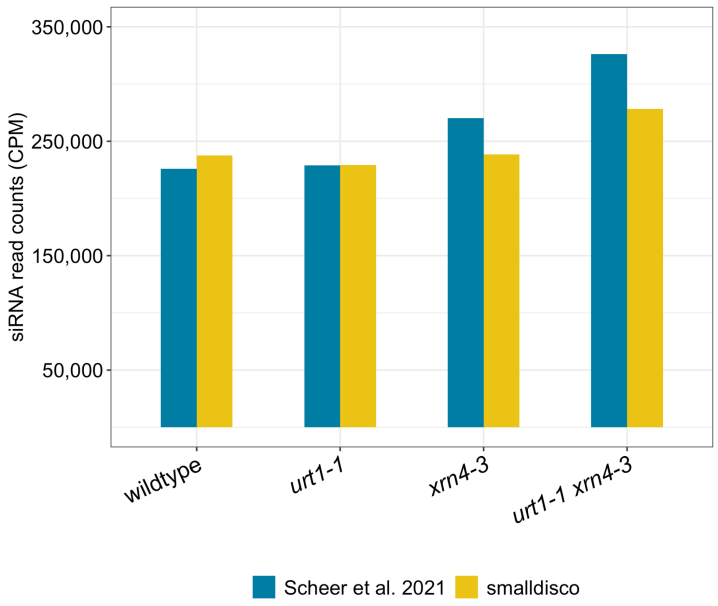

Figure 3 Identification of siRNA reads in small RNA sequencing datasets. SiRNA read counts are plotted as counts per million reads for small RNA sequencing datasets obtained for wildtype and three mutant genotypes. Blue bars indicate read counts reported by Scheer *et al.* (2021) and yellow bars indicate reads identified by smalldisco.

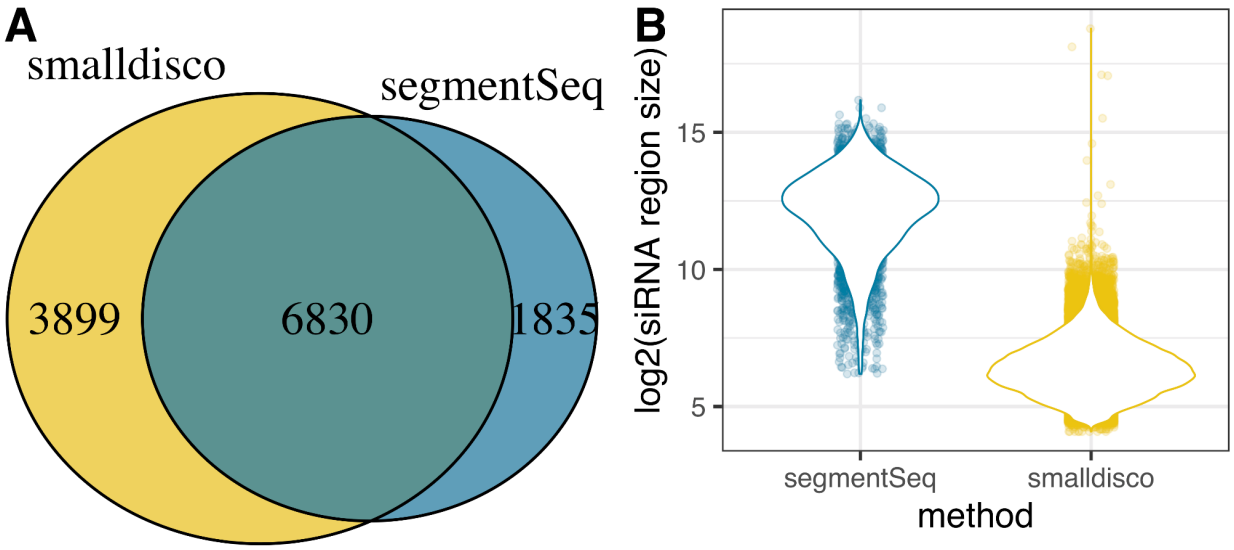

Figure 4 siRNA prediction comparison between smalldisco and segmentSeq. (A) Overlap between the number of sRNA target genes predicted using smalldisco and segmentSeq. Note that segmentSeq is not specific to siRNAs, but presumably most sRNAs derived from coding sequences are likely to be siRNAs. (B) The two methods differ substantially in the size of their predicted siRNA region, with smalldisco having an average size of 124.3 bp (±7.4 bp SE) and segmentSeq having an average size of 6.7kb (±68.9 bp SE).

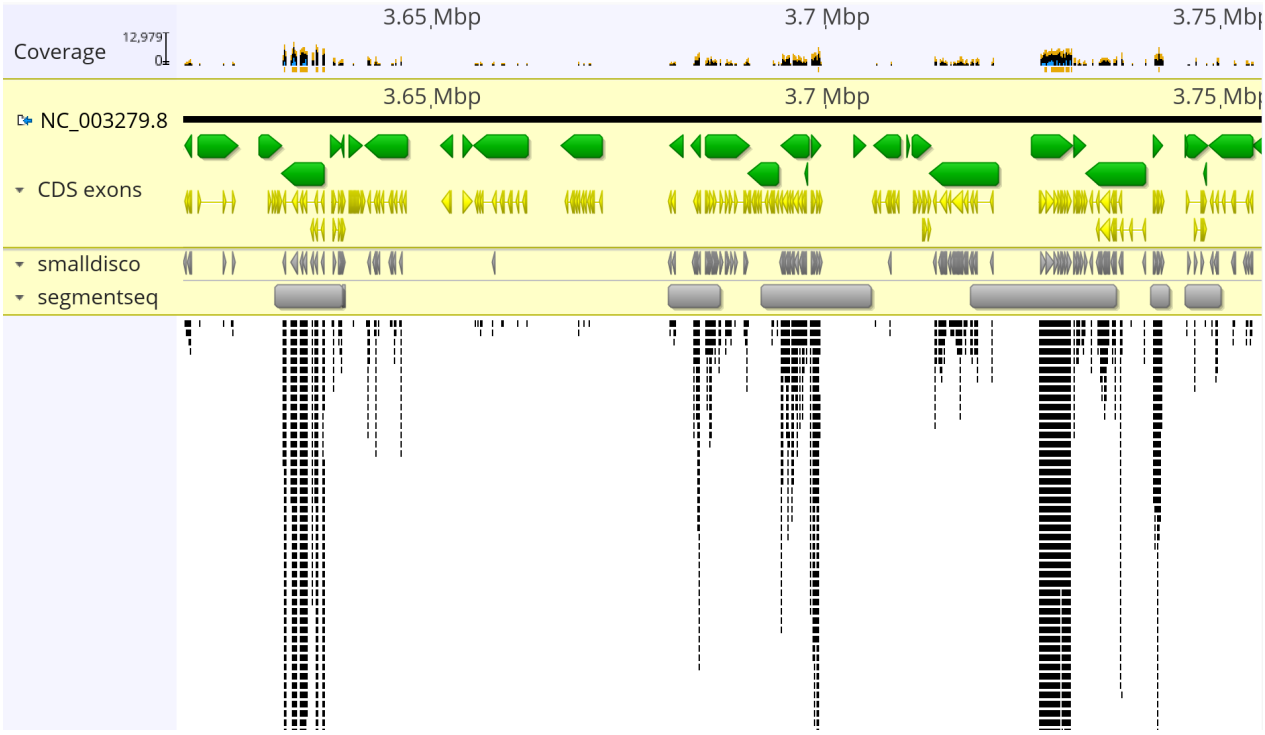

Figure S1 Examples of siRNAs identified by smalldisco and segmentSeq that map to one region of the genome. Image is a screen shot from the Geneious genome browser. Green bars indicate the positions of transcribed genes, with the pointed end indicating direction of transcription. Yellow bars indicate coding exons. Gray bars indicate siRNAs as identified by smalldisco (top) and segmentSeq (bottom). Stacked black lines represent siRNA-derived read pileups.

Table S1: List of NCBI accession numbers used in this study.

Paper	BioProject	Run(s)	Sample description	Figure
Xu et al. 2018	PRJNA447866	SRR6912753	3xFLAG::GFP::WAGO-4, 20°C	Figure 2A
Claycomb et al. 2009	PRJNA123485	SRR030720	small RNAs that co-immunoprecipitate with the argonaute CSR-1; CSR-1 immunoprecipitate from wild type, N2	Figure 2B
Scheer et al. 2021	PRJNA624174	SRR11521524	WT, replicate 1	Figure 3
		SRR11521526	urt1-1, replicate 1	
		SRR11521528	xrn4-3, replicate 1	
		SRR11521530	urt1-1 xrn4-3, replicate 1	
		SRR11521532	WT, replicate 2	
		SRR11521534	urt1-1, replicate 2	
		SRR11521536	xrn4-3, replicate 2	
		SRR11521538	urt1-1 xrn4-3, replicate 2	
Davis et al. 2018	PRJNA398116	SRR5931601	wildtype at 20C small RNA replicate 1	Figure 4
		SRR5931602	wildtype at 20C small RNA replicate 2	
		SRR5931603	wildtype at 20C small RNA replicate 3	
		SRR5931604	wildtype at 25C small RNA replicate 1	
		SRR5931605	wildtype at 25C small RNA replicate 2	
		SRR5931606	wildtype at 25C small RNA replicate 3	